Desenvolvimento de aplicações para Deficientes Visuais: Uma Discussão sobre Ferramentas para Incorporação da Tecnologia de voz ao VoiceProxy

Andréa dos Santos
Halisson Fabrício de Carvalho França
Ítalo Herbert Santos e Gomes
Wander Glayson Fernandes Teixeira
Guido Lemos de Souza Filho

Projeto Natalnet
Grupo de Ensino à Distância
Universidade Federal do Rio Grande do Norte
Email: {andrea, halisson, ihsg, wander, guido}@dimap.ufrn.br

Aplicações: Ensino à Distância

1- Introdução

Dentre as várias aplicações da área de Ensino à Distância podemos destacar a Educação Especial, que visa desenvolver tecnologias de hardware e software adaptando-os para auxiliar no processo de aprendizagem de pessoas que não possuem seu desenvolvimento cognitivo normal, tais como os deficientes visuais.

Através da exploração dos recursos das novas tecnologias da informação é possível criar ambientes de aprendizagem visando o desenvolvimento cognitivo dos portadores de necessidades especiais.

Ao projetar software para Educação Especial há certos requisitos que devem ser analisados, para que os objetivos do projeto sejam plenamente atingidos. De acordo com a deficiência em causa, variam as aptidões e necessidades dos utilizadores. Assim, a interface de software educacional deverá ser projetada de forma a melhor responder às necessidades do usuário. Com relação aos deficientes visuais destacamos alguns requisitos que devem ser atendidos pela interface, tais como a utilização de sons para interação usuário-máquina e privilegiando o uso do teclado através de teclas de atalho, evitando mensagens visuais e interação através do mouse.

No que se refere a interface dos software destinados aos deficientes visuais, podemos considerar o processo de síntese de voz como sendo um dos mais relevantes. Um sistema sintetizador *Text-To-Speech* (TTS) é um sistema capaz de ler textos, do idioma para o qual foi desenvolvido, e produzir os respectivos fonemas (sons), sejam textos obtidos diretamente da digitação do operador na máquina ou através de um sistema de reconhecimento ótico de caracteres denominado OCR (Recognition Character Optical).

No contexto deste trabalho, objetiva-se apresentar uma breve descrição das características e funcionalidade de dois sistemas para síntese de voz, o 'The Festi-

val Speech Synthesis System" e "IBM ViaVoice TTS SDK", apresentados como ferramentas para incorporação de tecnologia de voz no desenvolvimento de aplicações. Estas ferramentas foram analisadas como parte do desenvolvimento do VoiceProxy, um sistema em desenvolvimento no contexto do projeto NatalNet, cujo objetivo é sintetizar áudio a partir do processamento de páginas HTML.

2 - The Festival Speech Synthesis System

O Festival é tido como um sistema de síntese de voz para pelo menos três níveis de usuários. No primeiro nível, ele é destinado para aqueles usuários que simplesmente querem uma alta qualidade de voz de textos arbitrários com o mínimo de esforço. No segundo, ele é destinado para aqueles que estão desenvolvendo sistemas de idioma e desejam incluir saída sintetizada. Neste caso, é necessário uma certa quantidade de padronização é desejada, assim como vozes diferentes, phrasing específico, e etc. O terceiro nível consiste em desenvolver e testar novos métodos de síntese.

A filosofia adotada por sistemas como o Festival permite a adição e teste de novos módulos voz sem a necessidade de gastar esforços significativos para construir um sistema inteiro ou adaptar um já existente.

Existe uma outro aspecto do Festival que o torna mais útil do que um simples ambiente para introdução de novas técnicas de síntese. Ele é um sistema TTS inteiramente apropriado para ser utilizado em outros projetos que necessitem de saída de voz.

No Festival, nós podemos identificar três partes básicas do processo TTS, a fase de *Text analysis, Linguistic analysis* e *Waveform generation.*

A fase da *text analysis* tem como propósito colocar e organizar as orações em uma lista de gerenciamento de palavras, identificar números, abreviações e acrônimos transformando-os em texto cheio (ex. Sr.→ Senhor) quando necessário, utilizando uma gramática regular como base para solucionar alguns problemas; determinar a classe de cada palavra, individualmente, analisando a ortografia das mesmas e organizando uma lista de categorias e fazer a flexão e a derivação das palavras, quando necessário, decompondo-as em unidades gramaticais elementares através da análise de suas raízes léxicas e seus afixos (prefixos e sufixos); analisar as palavras observando o contexto em que elas estão inseridas, ou seja, analisando a palavra em questão associada aos seus vizinhos, possibilitando assim uma melhor identificação e diminuição da lista de categorias.

A fase da *linguistic analysis* é responsável pelo gerenciamento e produção da prosódia utilizada na geração dos sons. A prosódia se refere a certas propriedades de sinais da fala que estão relacionadas às mudanças de entonação de voz, sonoridade e duração do som das sílabas. A prosódia influi diretamente na comunicação por voz e tem uma função bastante específica nesse tipo de comunicação.

A fase de *waveform generation* é responsável pelo controle dinâmico das articulações e controle da freqüência vibratória das dobras vocais, que possibilitam a produção de sinais digitais exigidos.

O Festival está em constante desenvolvimento e pretende incluir diversos outros módulos. Aperfeiçoamentos já estão sendo considerados em vários estágios de implementação, como técnicas podemos citar síntese baseada em seleção, especificação léxica independente do dialeto, dentre outras.

3 - IBM ViaVoice TTS SDK

O IBM ViaVoice TTS SDK permite a incorporação das funcionalidades da tecnologia texto para fala no desenvolvimento de aplicações. Este SDK permite aos desenvolvedores a escolha entre duas APIs distintas: "Eloquence Command Interface" (ECI) e "Microsoft Speech Application Programming Interface" (SAPI). O IBM ViaVoice TTS SDK, juntamente com o IBM ViaVoice TTS RunTime, fornecem todos os softwares e arquivos de suporte para as duas APIs. ECI é uma API proprietária e independente de plataforma, que permite acesso direto a toda a funcionalidade do IBM ViaVoice TTS. Como características desta API destacam-se o seu suporte a diversos sistemas operacionais, padronização da saída de voz através de chamadas de funções e de anotações textuais, além de não utilizar o Registro do Windows para localização de componentes, evitando modificações acidentais de instalações por outras aplicações. SAPI é a API padrão da Microsoft, sendo suportada somente em sistemas Windows. Esta API fornece compatibilidade com padrões como ActiveX, COM, DCOM, MSAgent, e também permite padronização da saída de voz por meio de chamadas de funções e marcações de texto SAPI.

O IBM ViaVoice TTS SDK processa uma grande variedade de entrada textual, incluindo abreviaturas, acrônimos e números, e as pronuncia com fala de alta qualidade e uma entonação natural. Além disso, é possível personalizá-lo de diversas maneiras. É possível inserir marcações especiais no texto para alterar características de voz, ajustar entonação de sentenças e escolher modos de interpretação de texto e números. Pode-se utilizar ortografias fonéticas para especificar a pronúncia de uma palavra, permitindo armazenar estas pronúncias em um dos dicionários do usuário.

As marcações são códigos especiais que podem ser inseridos no texto para fazer o mecanismo de texto para fala se comportar de determinadas maneiras. As marcações controlam atributos como características da voz, ênfase de palavras, interpretação de números entre outros. Por exemplo: \Spd=282\ (fale 282 palavras por minuto).

Uma anotação tem a mesma função que uma marcação: é um código especial que pode ser colocado no texto de entrada para personalizar a saída. A maioria das anotações têm marcações equivalentes. As anotações devem ser utilizadas no lugar de marcações dentro da tradução de uma entrada do Dicionário de Palavras Especiais, descrito posteriormente.

O IBM ViaVoice TTS SDK fornece pelo menos cinco vozes predefinidas para cada idioma e cada uma tem uma marcação de voz correspondente que pode ser inserida no texto. Vozes individuais derivam sua exclusividade de diversos fatores físicos. Além disso, a voz de um indivíduo pode assumir diferentes qualidades em horas diferentes, dependendo de coisas como estado de espírito e circunstância. Estes atributos, tais como Trato vocal, Linha de Base de Tom, Tamanho da Cabeça, Rouquidão, Respiração, Flutuação de Tom, Velocidade e Volume, podem ser modificados com um conjunto de marcações de características de voz.

Uma Representação Fonética Simbólica (SPR) é a ortografia fonética que representa o som da palavra, como estes sons são divididos em sílabas e quais sílabas recebem ênfase, sendo o mecanismo utilizado pelo IBM ViaVoice TTS para representar as pronúncias de uma palavra. SPRs podem ser utilizadas quando as regras normais de letra para som não produzirem a pronúncia correta, para palavras específicas nos dicionários do usuário para que a pronúncia se aplique sempre que

uma determinada cadeia for encontrada, ou pode-se digitar SPRs no próprio texto de entrada, colocando-os na marcação \xSPR\.

O IBM ViaVoice TTS permite que se especifique pronúncias explícitas para palavras, abreviaturas e acrônimos, evitando a aplicação das regras normais de letra para som. Uma maneira para fazer isto é digitando uma marcação SPR diretamente no texto de entrada, como citado anteriormente. Uma maneira mais permanente é digitar a palavra (a cadeia de entrada ou ”chave”) e a pronúncia que se deseja (a saída ou ”tradução”) em um dos dicionários do usuário: Dicionário de Palavras Especiais, Dicionário de Abreviaturas, Dicionário de Radicais. O Dicionário de Palavras Especiais pode ser utilizado para: cadeias que se traduzem em mais de uma palavra, acrônimos (desde que a chave não contenha pontos na última posição), um endereço de e-mail, cadeias que contêm dígitos ou outros símbolos não-literais (que não são permitidos em outros dicionários), cadeias que requerem traduções com anotações ou SPRs. O Dicionário de Abreviaturas é utilizado para abreviaturas (com e sem pontos) que não requeiram a utilização de anotações em sua tradução. O Dicionário de Radicais é utilizado para palavras comuns como substantivos, verbos ou adjetivos e para nomes próprios. A propriedade distinta do Dicionário de Radicais é que ele armazena apenas a forma radical de uma palavra; todas as outras formas da palavra serão pronunciadas automaticamente da mesma maneira.

4 - Conclusões

Como descrito acima, o Festival pode ser utilizado tanto como um simples sistema para síntese de voz, como um completo ambiente de desenvolvimento, podendo vir a incorporar novos métodos de síntese. Já o IBM ViaVoice TTS SDK fornece aos programadores as ferramentas necessárias para o desenvolvimento de aplicações que incorporam a tecnologia de voz, incluindo um conjunto de APIs e utilitários que permitem ao desenvolvedor grande capacidade de padronização e gerenciamento do processo de síntese de voz acessado por uma aplicação.

No contexto do projeto NatalNet (www.natalnet.br), testamos as ferramentas descritas nas Seções 2 e 3 com o intuito de selecionar a que iremos utilizar para implementar um sistema leitor de páginas HTML Nosso objetivo é implementar um serviço que sintetiza áudio a partir de do processamento de páginas HTML. Uma vez pronto, o sistema permitirá que deficientes visuais naveguem através da Internet escutando o conteúdo das páginas.